DOCUMENT RESUME

ED 434 149 TM 030 113

AUTHOR	van der Linden, Wim J.
TITLE	Adaptive Testing with Equated Number-Correct Scoring. Research Report 99-02.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE	1999-00-00
NOTE	31p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Ability; *Adaptive Testing; Algorithms; Computer Assisted Testing; *Equated Scores; Estimation (Mathematics); Item Response Theory; Test Format
IDENTIFIERS	Law School Admission Test; *Number Right Scoring

ABSTRACT

A constrained computerized adaptive testing (CAT) algorithm is presented that automatically equates the number-correct scores on adaptive tests. The algorithm can be used to equate number-correct scores across different administrations of the same adaptive test as well as to an external reference test. The constraints are derived from a set of conditions on item response functions that guarantees the observed number-correct score distributions on two forms to be identical (W. van der Linden and R. Luecht, 1998). An item pool from the Law School Admission Test is used to compare the results of the algorithm with those for traditional observed-score equating of ability estimates to number-correct scores as well as the transformation to predicted number-correct scores through the test characteristic function. The effects of the constraints on the statistical properties of the ability estimator are examined. (Contains 18 references, 4 figures, and a list of University of Twente research reports.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM

ED 434 149

Adaptive Testing with Equated Number-Right Correct Scoring

Research Report 99-02

Wim J. van der Linden

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Nelissen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM030113

BEST COPY AVAILABLE

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

Department of Educational M[redacted]ata Analysis

University of Twente

2

Adaptive Testing with Equated Number-Correct Scoring

Wim J. van der Linden

Adaptive Testing with Equated Number-Correct Scoring

Wim J. van der Linden

University of Twente

Send request for information to: W.J. van der Linden, Department of Educational Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. Email: vanderlinden@edte.utwente.nl

Abstract

A constrained CAT algorithm is presented that automatically equates the number-correct scores on adaptive tests. The algorithm can be used to equate number-correct scores across different administrations of the same adaptive test as well as to an external reference test. The constraints are derived from a set of conditions on item response functions that guarantees the observed number-correct score distributions on two test forms to be identical (van der Linden & Luecht, 1998). An item pool from the Law School Admission Test is used to compare the results of the algorithm with those for traditional observed-score equating of ability estimates to number-correct scores as well as the transformation to predicted number-correct scores through the test characteristic function. The effects of the constraints on the statistical properties of the ability estimator are examined.

Key words: Computerized Adaptive Testing, Item Response Theory, Observed-Score Equating, Optimal test Assembly, 0-1 Linear Programming

Adaptive Testing with Equated Number-Correct Scoring

At least three practical cases exists in which a method to equate number-correct scores on a computerized adaptive test (CAT) would be welcome. First, to accommodate preferences among its examinees, a testing program may offer them the choice between an adaptive and a paper-and-pencil version of the same test. This choice, available, for example, for the Armed Services Vocational Aptitude Battery (ASVAB) (Segall, 1997), is only fair if examinees can be guaranteed comparable scores on the two versions of the test. Second, to enhance the interpretability of its scores, a CAT program may report its scores as predicted number-correct scores on a released paper-and-pencil version of the test. This practice is followed, for example, in the Scholastic Assessment Test (SAT) (Lawrence & Feigenbaum, 1997). Third, though it makes perfect sense to correct test scores for the properties of the items placing all examinees on a common scale, and IRT-based CAT capitalizes on this feature, it is a well-known experience that the majority of the examinees who take a CAT still tend to focus on their number of items correct. These examinees may get confused if they answer more items correct but get a lower score than examinees with fewer items correct. Adaptive testing with number-correct scores of different examinees automatically equated to each other would help to level an important psychological barrier to accepting CAT.

One approach to solving the problem of score comparability between a CAT and a paper-and-pencil test is to estimate empirically the transformation that places CAT ability estimates on the number-correct score scale of the paper-and-pencil test for a population of examinees using the technique of observed-score equating. For the ASVAB, equipercentile equating of ability estimates and observed scores in combination with a randomly equivalent-groups design has been used (Segall, 1997). The same technique, albeit in combination with a nonequivalent-group common-items design, was used in Lawrence and Feigenbaum (1997).

Another approach is followed, for example, in the SAT program where the ability

estimates of the examinee are used to predict their scores on a released version of the SAT. Let the released version of the test have items indexed by j=1,...,n with response functions defined by the three-parameter logistic (3PL) model:

$$P_j(\theta) \equiv c_j+(1-c_j)[1+\exp(-a_j(\theta-b_j))]^{-1}, \quad (1)$$

where $\theta\in(-\infty,\infty)$ is a parameter for the abilities of the examinees and $b_j\in(-\infty,\infty)$, $a_j\in[0,\infty)$ and $c_j\in[0,1]$ are the difficulty, discriminating power and guessing parameter of item j, respectively. The same model was used in the empirical example below. In addition, let $\hat{\theta}$ be the ability estimate of an examinee on the CAT. A prediction of the number-correct score on the released version of the test, Y, for this examinee is obtained from the test characteristic function as:

$$\hat{\tau}_Y \equiv \tau_Y(\hat{\theta}) = \sum_{j=1}^{n} P_j(\hat{\theta}). \quad (2)$$

For the SAT a modification of this transformation is used to correct for guessing (see below). Notice that Equation 2 is not an observed score but an estimated expected observed score (that is, estimated true score). Further, since the test characteristic function for the CAT is also known, Equation 2 in fact relates the true score associated with an ability estimate on the CAT to the true score on the released version of the test. Thus the practice of predicting scores on a released test form through its test characteristic function can be viewed as approximate IRT true-score equating. The technique of IRT true-score equating is explained in Kolen and Brennan (1995) and Lord (1990).

A third approach is followed by Stocking (1996; see also Yen, 1984). To deal with the complexity of IRT-based test scoring, she proposed to modify the likelihood equations such that the solution to this equations becomes a monotonic function of the number-correct score of the examinee. In combination with the transformation through the test

characteristic function in Equation 2, this modification produces estimated true number-correct scores on a reference test that have the same ranking as the observed number-correct scores on the CAT.

Though actual observed-score equating is guaranteed, a practical disadvantage of equipercentile equating of a CAT to a paper-and pencil version of the test is the need to run a separate empirical study prior to the introduction of the CAT program. Such studies typically involve quite an amount of time and resources and need to be repeated each time the item pool or any other specification of the CAT program is changed. A typical way to reduce the costs is to use smaller samples than actually required using a smoothing technique to allow for the sample size. However, such techniques are likely to induce an (unknown) bias in the estimated equating transformation. Other potential threats to the validity of observed-score equating are the difficulty to realize common test administration conditions throughout the study and the inability to deal with scores at the lower end of the scale which are likely to suffer from guessing.

As for the test-characteristic-function approach, transformations to scores on earlier released versions of the test as those in Equation 2 are obtained as a direct spinoff from regular item-pool calibration. Use of this transformation thus avoids the practical problems involved in actual equating studies. However, it is hard to claim score comparability of the transformed CAT scores and the observed scores on the reference test. If the true θ value in Equation 2 were used, the transformation would amount to exact IRT true-score equating. However, substituting an estimate of θ does not result in observed-score equating because the error distributions of the observed scores on the CAT are not identical to those of the θ estimates upon transformation by Equation 2.

Though the modified-likelihood approach is attractive in that neither an additional equating study is required, it has the disadvantage of introducing an estimator of θ that does not belong to any familiar class of estimators. Unlike the maximum-likelihood (ML) or the Bayesian estimators currently used in CAT, the estimator does not have known

(asymptotic) properties. In fact, the nonmonotonic relation of the estimator to the MLE for the 2PL and 3PL models shows that information in the data is lost. Though this property may be offset by choosing a somewhat longer test, the estimator is also inconsistent.

This paper proposes to solve the problem of score equating by imposing a set of constraints on the CAT item selection algorithm that automatically equates the number-correct scores on the adaptive test. The constraints are derived from a set of conditions on the response functions of the items that guarantees the observed number-correct score distributions on two test forms to be identical (van der Linden & Luecht, 1998). To impose the constraints on the item selection, the method of constrained CAT with shadow tests in van der Linden and Reese (1998) is used.

As the algorithm *selects* the items to have the same number-correct scores distribution as the scores on a reference test, no additional transformation or equating is needed. As a result, observed number-correct scores can be compared directly to each other. Thus it is no longer possible to answer more items on a CAT correct but receive a lower score than an examinee with fewer items correct. On the other hand, the method can only be used for fixed-length CAT and reference tests that have the same length as the CAT. Further, as in any other IRT-based method, the proposed method relies heavily on the assumption that the item pool has excellent fit to an item response model. Further evaluation of the method will be postponed until it has been described more extensively and some empirical results have been presented.

The next section reviews the conditions on the response functions that guarantee two test forms to have identical observed score distributions and show how these conditions can be implemented in a CAT algorithm. Then modifications of the method to deal with formula scoring test forms are discussed. The next section presents results from a empirical study in which the efficacy of the method for an item pool from the Law School Admission Test (LSAT). The study was conducted to compare the results of the method with those for traditional observed-score equating of ability estimates to number-

correct scores as well as those for the transformation using the test characteristic function. Another goal of the study was to assess the effects of the constraints on the statistical properties of the CAT ability estimator. An evaluation of the proposed algorithm as well as a few remaining practical issues are addressed in the final section of the paper.

CAT Algorithm for Equated Number-Correct Scores

Let X denote the observed score on a test with items i=1,...,n and Y the observed score on another test with items j=1,...,n. van der Linden and Luecht (1998) prove that, for any ability distribution $h(\theta)$, the distributions of X and Y are identical if and only if

$$\sum_{i=1}^{n} P_i^r(\theta) = \sum_{j=1}^{n} P_j^r(\theta), \quad -\infty<\theta<\infty, \tag{3}$$

for r=1,...,n.

These conditions thus require that the sums of the first through the nth power of the response functions in the two test be equal. However, these authors also show that for $n\rightarrow\infty$ the conditions for r>2 become negligible. As a consequence, for realistic test lengths use of the first 2-3 conditions only already gives excellent approximations. For r=1, the condition in Equation 3 equates the true scores on the two test.

Applications to Adaptive Testing

For fixed tests and a population of examinees, the conditions in Equation 3 would require the sums of the powers of the response functions to be identical over the full range of θ values. However, if one of the tests is a CAT and each examinee gets an individual set of items from the pool, these sums then have to be identical *only for true* θ *value of the examinee*. For a population of CAT examinees, it follows that the marginal distributions of observed scores are always identical since their conditional distributions are.

As a generic term, the term "reference test" will be used to denote the test to which the CAT is equated. In view of the applications addressed in this paper, three different

types of reference tests are distinguished:

1. A currently used paper-and-pencil version of the same test.
2. A released earlier version of the test.
3. A dummy test with a conveniently selected set of response functions.

The first two cases have already been discussed extensively. Notice that in either of these cases it hold that, since each CAT administration is equated to the same reference test, *they are automatically equated mutually*.

The same principle can be exploited if no equating to an external reference test is intended. In this case, the response functions of a conveniently chosen dummy test can be used to formulate the right-hand side of the conditions in Equation 3. One possibility is to use the response functions of an actual CAT for this purpose, for example, a CAT actually administered to an average examinee. However, the response functions do not even need to belong to existing items. Any set of response functions with convenient parameter values will do.

Constrained Adaptive Testing

The idea is to impose the conditions in Equation 3 on the item selection in the CAT for enough powers of the response functions to create identical observed-score distributions for the examinees. An efficient way to implement these conditions is through the method of constrained CAT with shadow tests in van der Linden and Reese (1998).

At each step, items are selected not directly form the pool but from a complete optimal test selected from the pool ("shadow test"). The shadow test for the administration of the kth item has to meet the following specifications:

1. It should have maximum information at the current ability estimate.
2. Its length should be equal to the number of items in the (fixed-length) CAT.
3. It should meet all constraints on the CAT.
4. It should contain the k-1 items already administered.

From the items in the shadow test not yet administered, the one with maximum

information at the current ability estimate is selected as the kth item in the CAT. The procedure is repeated n times. Thus, the first item is selected form the full shadow test; for the last item only one free item in the shadow test is left.

Since each shadow test has to meet all constraints, future item selection always remains feasible with respect to the set of constraints to be imposed on the CAT. Also, because both the shadow tests and the individual items are selected to have maximum information, the CAT tends to be maximally informative too. Further details of this method of constrained CAT with shadow tests can be found in van der Linden and Reese (1998). An application of the method to the problem of controlling differential speed in CAT is given in van der Linden, Scrams and Schnipke (1999).

As the conditions in Equation 3 are linear in the items, shadow tests can be selected using the technique of 0-1 linear programming (LP). An introduction to 0-1 LP test assembly is given in van der Linden (1998a).

Model for Selection of Shadow Tests

A model for assembling shadow tests that guarantees equated number-correct is formulated. Let the items in CAT pool be denoted by index i=1,...,I. In addition, the items in the CAT are denoted by index k=1,...,n. Thus, i_k is the index of the item in the pool administered as the kth item in the CAT. The set indices of the first k-1 items in the CAT is thus $S_{k-1} \equiv \{i_1,...,i_{k-1}\}$. As before, the items in the reference test are denoted as j=1,...,n. Further, $\hat{\theta}^{(k-1)}$ is the estimated value of θ after k-1 items have been administered.

To formulate the model, binary variables x_i are used to denoted whether (x_i=1) or not (x_i=0) item i is selected in the shadow test. The model for the selection of the kth shadow test is:

$$\text{maximize} \quad \sum_{i=1}^{I} I_i(\hat{\theta}_{k-1})x_i \tag{4}$$

subject to

$$\sum_{i=1}^{I} P_i^r(\hat{\theta}_{k-1})x_i - \sum_{j=1}^{n} P_j^r(\hat{\theta}_{k-1}) \leq c, \; r=1,...,R \tag{5}$$

$$\sum_{i=1}^{I} P_i^r(\hat{\theta}_{k-1})x_i - \sum_{j=1}^{n} P_j^r(\hat{\theta}_{k-1}) \geq -c, \; r=1,...,R \tag{6}$$

$$\sum_{i=1}^{I} x_i = n, \tag{7}$$

$$\sum_{i \in S_{k-1}} x_i = k-1, \tag{8}$$

$$x_i \in \{0,1\}, \; i=1,...,I. \tag{9}$$

The objective function in Equation 4 maximizes the information in the shadow test at the current ability estimate $\hat{\theta}_{k-1}$. The constraints in Equation 5 and 6 require the differences between the sums of the first R powers of the probabilities of success for the items in the paper and shadow test to be in identical up to c, where c is a tolerance parameter with a small value chosen by the CAT administrator. If necessary, different values of c for different powers of the probabilities of success can be chosen. The test length is set equal to n by the constraint in Equation 7 whereas the constraint in Equation 8 requires the previous k-1 items to be in the shadow test for the kth item. The constraints in Equation 8 define the range of the decision variables.

The tolerance factor in Equation 5 and 6 is introduced for technical reasons only; imposing an exact equality is likely to lead to infeasibility of the problem. Asymptotic consequences of using $\hat{\theta}$ rather than θ in Equation 5 and 6 are discussed below.

Models as in Equations 4 through 9 can be solved for optimal values of their decision variables using one of the algorithms or heuristics available in general LP software, for example, CPLEX (ILOG, 1998), or the test assembly package ConTEST

(Timminga, van der Linden & Schweizer, 1997).

Extensions and Special Cases

As identity of distributions is maintained under identical transformation of their variables, the conditions in Equation 3 guarantee equating of any monotonic function of number-correct scores. A transformation often used with multiple-choice items is formula scoring to correct for guessing:

$$\frac{aX - n}{a-1}, \tag{10}$$

where X is the number-correct score on the test and a is the (common) number of alternatives per item (Lord & Novick, 1968, eq. 14.3.4). Since the relation in Equation 10 is linear in X, formula scores are automatically equated under the conditions in Equation 3.

An alternative to the transformation through the test characteristic function in Equation 2 is Lord's (1980, eq. 15.6) true formula score with θ replaced by its estimated value:

$$\frac{a\sum_{j=1}^{n} P_j(\hat{\theta}) - n}{a-1} \tag{11}$$

This transformation was used as an analogue to the observed-score transformation in Equation 10 by Lawrence and Feigenbaum (1997).

Notice that the transformation in Equation 11 with true θ values would equate true formula scores. For estimates of θ, the transformation equates neither true nor observed formula scores. However, true formula score equating is possible by selecting the shadow test such that the formula in Equation 11 is equated to the one for the reference test. Requiring these formulas to be equal up to a tolerance factor c gives the following constraints:

$$\sum_{i=1}^{I} P_i(\hat{\theta}_{k-1})x_i - \sum_{j=1}^{n} P_j(\hat{\theta}_{k-1}) \leq (a-1)c/a, \quad (12)$$

$$\sum_{i=1}^{I} P_i(\hat{\theta}_{k-1})x_i - \sum_{j=1}^{n} P_j(\hat{\theta}_{k-1}) \geq -(a-1)c/a. \quad (13)$$

These constraints are identical to those in Equation 5 and 6, except for a rescaling of the tolerance factor c. If exact equality were required, that is, if c=0, then, as expected, true score and true formula score equating would yield the same results.

Discussion

It is emphasized that the constraints in Equation 5 and 6 equate *sums* of powers of success probabilities and no powers of individual probabilities. Hence, these quantities can compensate each other across items. Thus, when selecting the items, space for optimization is present. In fact, as follows from Proposition 4 in van der Linden and Luecht (1998), equating of the individual probabilities in the CAT to those in the reference test would be implied if all n conditions in Equation 3 were imposed. Since CATs typically have 25-30 items and only the sums of the first 2-3 powers have to be equated, the space for optimization is expected to be generally substantial.

Notice that the conditions in Equation 3 are formulated for true θ values whereas they are implemented for the current estimate of θ in the constraints in Equation 4 and 5. However, as shown in Chang and Ying (in press), for an infinitely large item pool, the maximum-likelihood estimator of θ in a CAT with maximum-information item selection is strongly consistent for the 1PL model. For the 2PL model, strong consistency holds provided realistic bounds on the values of the discrimination parameters are met. For the 3PL model, the same results hold again provided an additional realistic bound on the guessing parameter is met and the likelihood equations have no multiple solutions. As the conditions in Equation 3 are based on continuous functions of θ, it follows from Slutsky's theorem for strong convergence (e.g., Ferguson, 1996) that the differences in the left-hand sides of Equation 5 and 6 also converge to their true-θ equivalents. These results are

expected to be closely approximated for well-designed finite item pools CAT with MLE and maximum-information item selection. The presence of the constraints in Equation 5 and 6 in the CAT algorithm only amounts to a reduction of the effective size of the item pool. However, since the reduction is to the most informative subset of items, the effect is expected to be only a slightly slower rate of convergence.

Thus, typically, a CAT session is expected to take a course in which the first items, selected at a value of $\hat{\theta}$ off target, tend to cumulate partial sums of powers of success probabilities at the true θ value of the examinee that do not match those of the reference test. Because $\hat{\theta}$ converges to its true value, contributions by the items later in the process will tend to compensate for earlier contributions and the differences between the sums in Equation 5 and 6 converge to their true equivalents.

Empirical Example

The LSAT item pool consisted of 753 items all calibrated using the 3PL model in Equation 1. The pool was assembled from previously administered paper-and-pencil versions of the test. An arbitrary form was identified to select randomly a set of reference tests of n=10(10)50 to which the CATs had to be equated. The set of reference tests was nested.

The following conditions were simulated:

1. Unconstrained CATs of n=10(10)50 items.
2. Unconstrained CATs of n=10(10)50 items with true number-correct scores on the reference test estimated through its test characteristic function (Equation 2).
3. Constrained CATs of n=10(10)50 items with shadow tests selected for R=1(1)4 (Equations 4 through 9).

The number of replications for each CAT was equal to 8,000 for Condition 1 and 2 and 1,800 for each level of Condition 3.

For each condition, after 10(10)50 items the following data were collected:

1. Observed number-correct scores (estimated true number-correct scores in Condition 2).
2. Estimated bias in $\hat{\theta}$.
3. Estimated mean squared error (MSE) in $\hat{\theta}$.

The target distributions to which the observed number-correct score distribution in Condition 3 and the estimated true number-correct score distributions in Condition 2 were equated were the observed number-correct score distributions for the references tests. The target distributions were generated from the item parameters for the reference tests for $\theta \sim N(0,1)$ using the algorithm for the generalized binomial distribution described in Lord and Wingersky (1984). The length of the CATs was varied to examine the speed of convergence of the observed-score distributions to their targets. Finally, the effects of the constraints in Equation 5 and 6 on the statistical properties of $\hat{\theta}$ were assessed. To do so, the estimated bias and MSE functions of this estimator for Condition 1 and 3 were compared.

Adaptive tests were simulated according to a procedure for Bayesian initialization of the θ estimator in van der Linden (1999). The true θ values of the simulees were randomly drawn form N(0,1). Then, given the value of θ, a value on a background variable X was sampled. The bivariate distribution of θ and X was assumed to be standard normal, with $\rho_{\theta X}=.\bar{6}0$. (Nearly the same correlation was found in the empirical example in van der Linden, 1999). The initial estimate of θ was the regressed value of θ on X. The first shadow test was assembled to be maximally informative at the estimate. Next estimates were obtained using the same EAP estimator. This estimator is known to perform generally well with a smaller MSE than the maximum-likelihood and a slight inward bias (van der Linden, 1998b). Also, unlike the maximum-likelihood estimator, it always exists.

Implementation of Algorithm

Trial runs with the algorithm showed an occasional case of infeasibility if the

values of c in the constraints in Equation 5 and 6 were chosen too tight. In such cases, the first items administered appeared to be highly informative at θ estimates largely off target. As a result, the sum of the response functions for the first part of the CAT became steep at the wrong θ value, and it became difficult for the full test to meet the constraints in Equation 5 and 6 for small c values at other values of θ.

To deal with such cases, the algorithm was implemented as follows: First, the response functions in the CAT were constrained to satisfy Equation and 6 not only at the current $\hat{\theta}$ value but also at values slightly lower and higher than $\hat{\theta}$. This measure was introduced to make the algorithm more robust with respect to θ estimates too much off target. In this example, the additional constraints were formulated at $\hat{\theta}-.5$ and $\hat{\theta}+.5$. Second, the algorithm was started with a small value of c. As soon as a case of infeasibility was met, the additional constraints were removed if they caused the infeasibility or the value of c was slightly increased. In this example, the algorithm was started with c=.5 and the increase was set at .2.

The simulations were run on a PC with Pentium Pro/166MHz processor. The LP-models for the shadow tests were solved through calls to the CPLEX 6.0 software (ILOG, 1998). Because solutions to 0-1 LP models for test assembly are calculated iteratively, good starting values are necessary. For this purpose, a shadow test optimal at θ=0 was calculated prior to the simulations which was used as starting solution in each CPLEX run in the simulation. As a result, all solutions in the simulation were obtained in 6-8 seconds.

Results

The observed number-correct score distributions for unconstrained CAT (Condition 1) and the constrained CAT versions (Condition 3) are plotted against their target distributions in Figure 1. The distributions for unconstrained CAT show their typical

[Figure 1 about here]

peaked form. The target distribution on the reference test was much wider. For n=10, all distributions for constrained CAT were between those for unconstrained CAT and the

target distribution but much closer to the latter. With increasing test length, the target distribution was approximated better and better. For n=30, the approximation was already good, whereas for n=50 no systematic differences between the distributions and the target were left. The value of R did not seem to have an impact, with the exception of the case of n=10 for which the distribution for R=2 was slightly superior to those for R=1, 3, 4, and 5.

In Figure 2, the distributions of the true number-correct scores on the reference test estimated through its test characteristic function (Condition 2) are shown along with the observed-score distributions under unconstrained CAT (Condition 1) and

[Figure 2 about here]

the target distribution on the reference test. For all test lengths, the distributions of estimated true number-correct scores had smaller variability than their target distribution. The approximation improved considerably with increasing test length, though. Further, a typical distortion of the lower tail of the distribution was observed. Due to the lower asymptote of the test characteristic function introduced by the guessing parameter in the 3PL model, observed scores between the level corresponding with this asymptote are impossible.

In Figure 3, the estimated bias in the estimator of θ is plotted as a function of θ

[Figure 3 about here]

for the unconstrained (Condition 1) and constrained CAT versions (Condition 3). The general shape of these plots shows a well-known inward bias for the EAP estimator. Except for n=10, the presence of the constraints in Equation 5 and 6 did not appear to introduce any additional bias in the estimator. For n=10, an increased inward bias at the lower end of the θ scale and more variation for the various values of R at the upper end of the scale was found.

For the same CAT conditions, in Figure 4 the estimated MSE in the estimator is plotted as a function of θ. These plots show the price that has to be paid for the presence

[Figure 4 about here]

of the constraints in Equation 5 and 6 in the CAT algorithm. For all test lengths, the MSE for constrained CAT is systematically larger than for unconstrained CAT. The largest relative loss of efficiency occurred for n=10 at the lower end of the θ scale. For longer tests, the loss of efficiency decreased but was still too large to be ignored.

Conclusions

In the empirical example, the constrained CAT algorithm proposed in this paper did not show any unexpected behavior. For longer tests, it produced observed number-correct score distribution that did not differ systematically from the target on the reference test. Also, it did not introduce any systematic bias in the θ estimator. However, the estimator did loose some of the efficiency associated with regular, unconstrained CAT. Users of the algorithm should be prepared to accept this loss or compensate for it by accepting a longer test length. For shorter tests, the empirical example yielded observed number-correct score distributions that were more peaked than the distributions on the reference test. In particular for n=10, additional equating seems to be necessary. However, use of the algorithm is then still recommended because it minimizes the distortion of the number-correct scale involved in the additional equating. The relative sizes of the differences between the distributions for constrained and unconstrained CAT and their targets in the plots in Figure 1 show how large the gain can be.

For all test lengths, the algorithm outperformed the transformation through the test characteristic function in Equation 2, currently one of the standards of the testing industry. Also, it is reminded that these results were obtained for *actual* numbers of items correct in the CAT rather than a *post hoc* transformation of the θ estimator with an indirect direct relation to the response vectors. This property is believed to be the most practical feature of the CAT algorithm proposed in this paper.

References

Chang, H. & Ying, Z. (in press). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. The Annals of Statistics.

Ferguson, T. S. (1996). A course in large-sample theory. London: Chapman & Hall.

ILOG (1998). CPLEX 6.0 Documentation supplement [Computer software]. Incline Village, NV: ILOG, Inc.

Lawrence, I, & Feigenbaum, M. (1997). Linking scores for computer-adaptive and paper-and-pencil administrations of the SAT (Research Report 97-12). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". Applied Psychological Measurement, 8, 452-461.

Mislevy, R. J. & Wu, P.-K. (1988). Inferring examinee ability when some items are missing (Research Report 88-48). Princeton, NJ: Educational Testing Service.

Segall, D. O. (1997). Equating the CAT-ASVAB. In W. A. Sands, B. K. Waters. & J. R. McBride (Eds.), Computerized adaptive testing: From inquiry to operation (pp. 181-198). Washington, DC: American Psychological Association.

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. Journal of Educational and Behavioral Statistics, 21, 365-389.

Timminga, E., van der Linden, W. J., & Schweizer, D. A. (1997). ConTEST 2.0 Modules: A decision support system for item banking and optimal test assembly

[Computer program and manual]. Groningen, The Netherlands: iec ProGAMMA.

van der Linden, W. J. (1998a). Optimal assembly of psychological and educational tests. Applied Psychological Measurement, 22, 195-211.

van der Linden, W. J. (1998b). Bayesian item selection criteria for adaptive testing. Psychometrika, 63, 201-216.

van der Linden, W. J. (1999). A procedure for empirical initialization of the trait estimator in adaptive testing. Applied Psychological Measurement, 23, 21-29.

van der Linden, W. J. & Luecht, R. M. (1998). Observed-score equating as a test assembly problem. Psychometrika, 62, 401-418.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. Applied Psychological Measurement, 22, 259-270.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for speededness in computerized adaptive testing. Applied Psychological Measurement, 23. (In press)

Yen, W. M. (1984). Obtaining maximum-likelihood estimates from number-correct scores for the three-parameter logistic model. Journal of Educational Measurement, 21, 93-111.

Author Note

This study received funding from the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of LSAC. The author is indebted to Wim M.M. Tielen for his computational assistance. Send all correspondence to: W.J. van der Linden, Department of Educational Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. Email: vanderlinden@edte.utwente.nl.

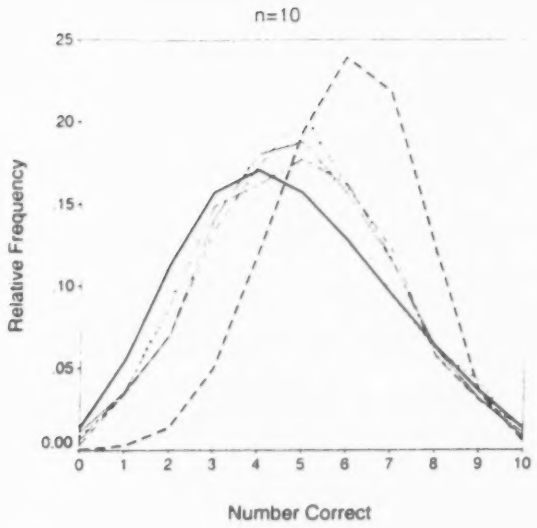
Figure Captions

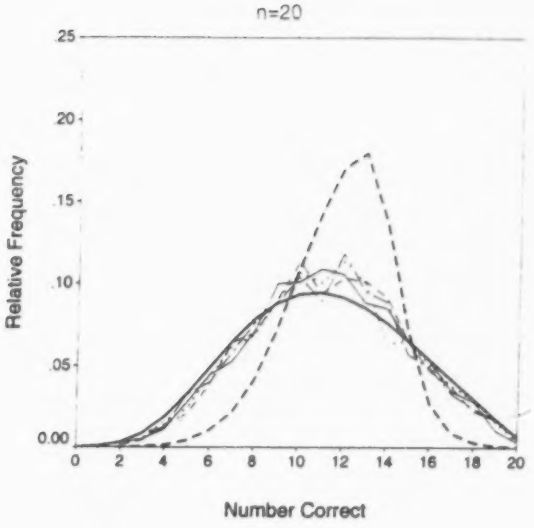
Figure 1. Distributions of observed number-correct score for unconstrained CAT (UCAT), reference test (Target) and constrained CAT (R=1,2,3,4) for n=10, 20, 30, 40, and 50.

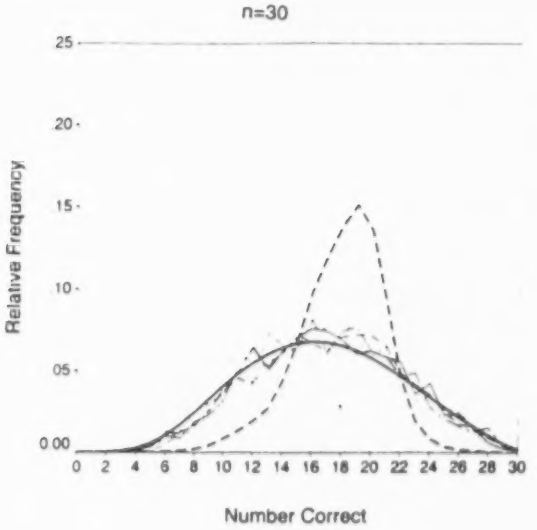
Figure 2. Distributions of observed number-correct score for unconstrained CAT (UCAT) and reference test (Target), and distributions of true number-correct scores on reference test estimated through its test characteristic function (TCC) for n=10, 20, 30, 40, and 50.

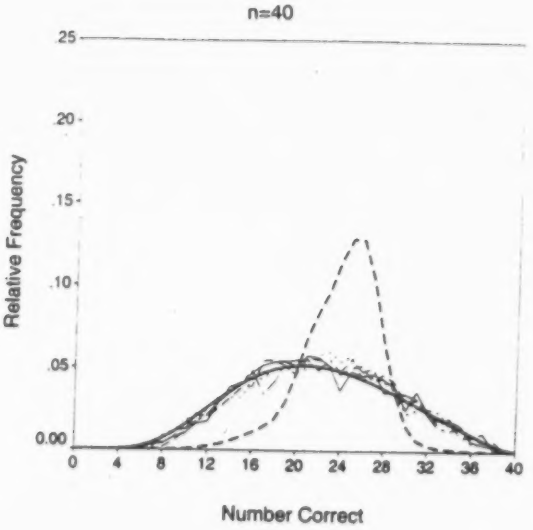
Figure 3. Bias functions for unconstrained CAT (UCAT) and constrained CAT (R=1,2,3,4) for n=10, 20, 30, 40, 50.

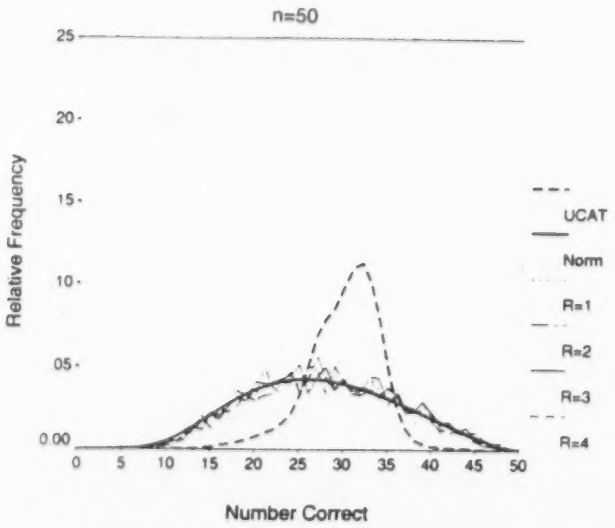
Figure 4. MSE functions for unconstrained CAT (UCAT) and constrained CAT (R=1,2,3,4) for n=10, 20, 30, 40, 50.



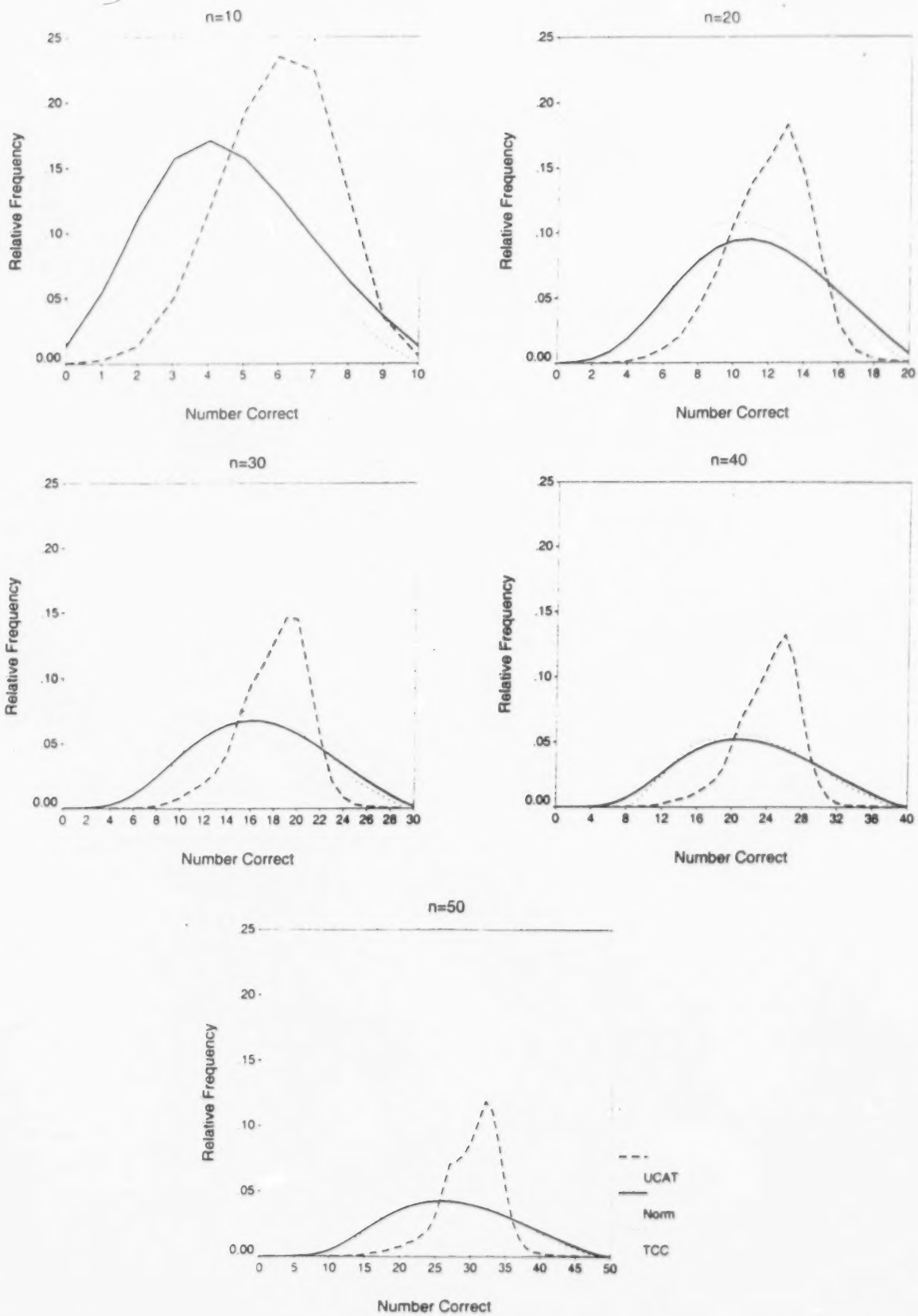




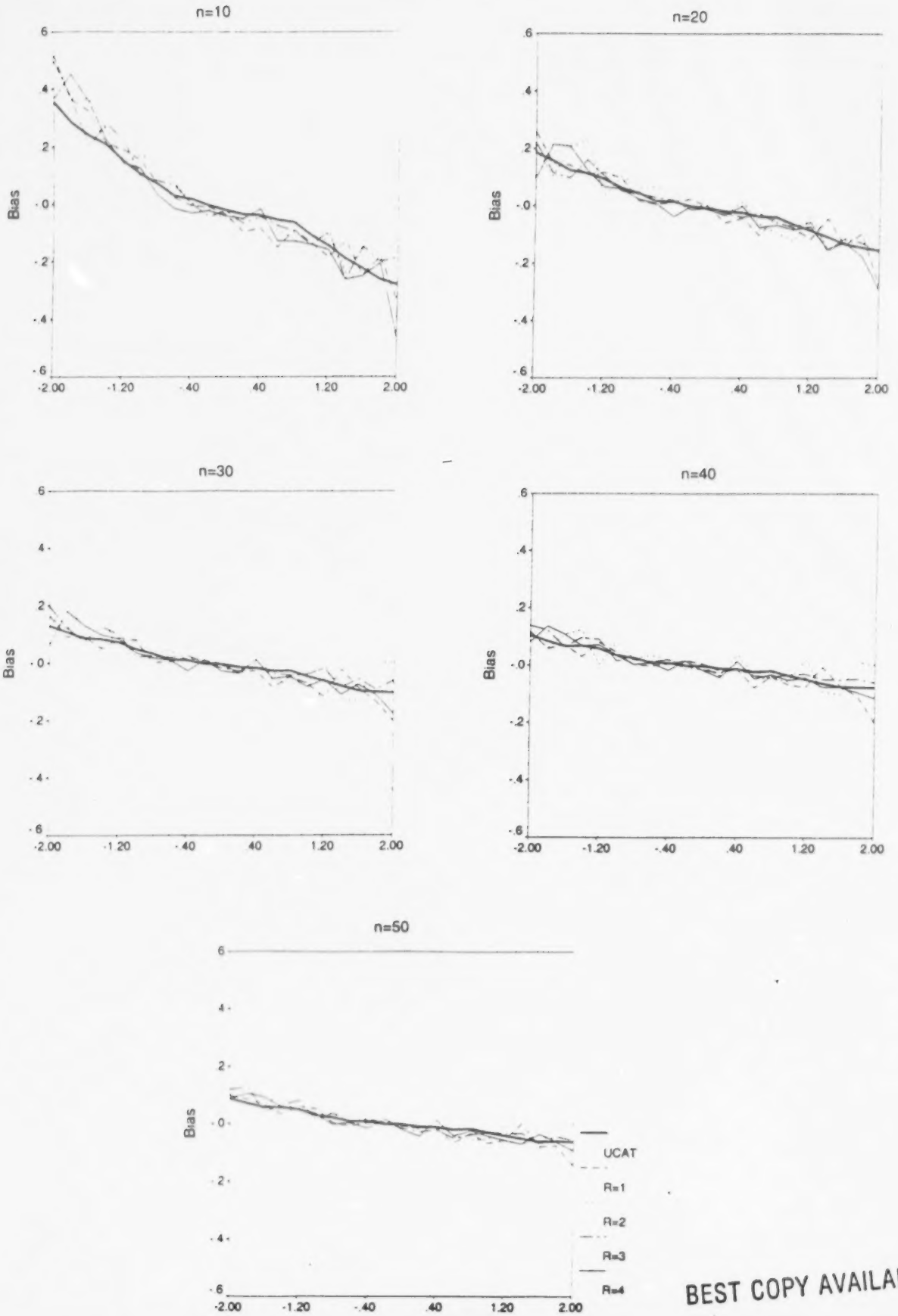




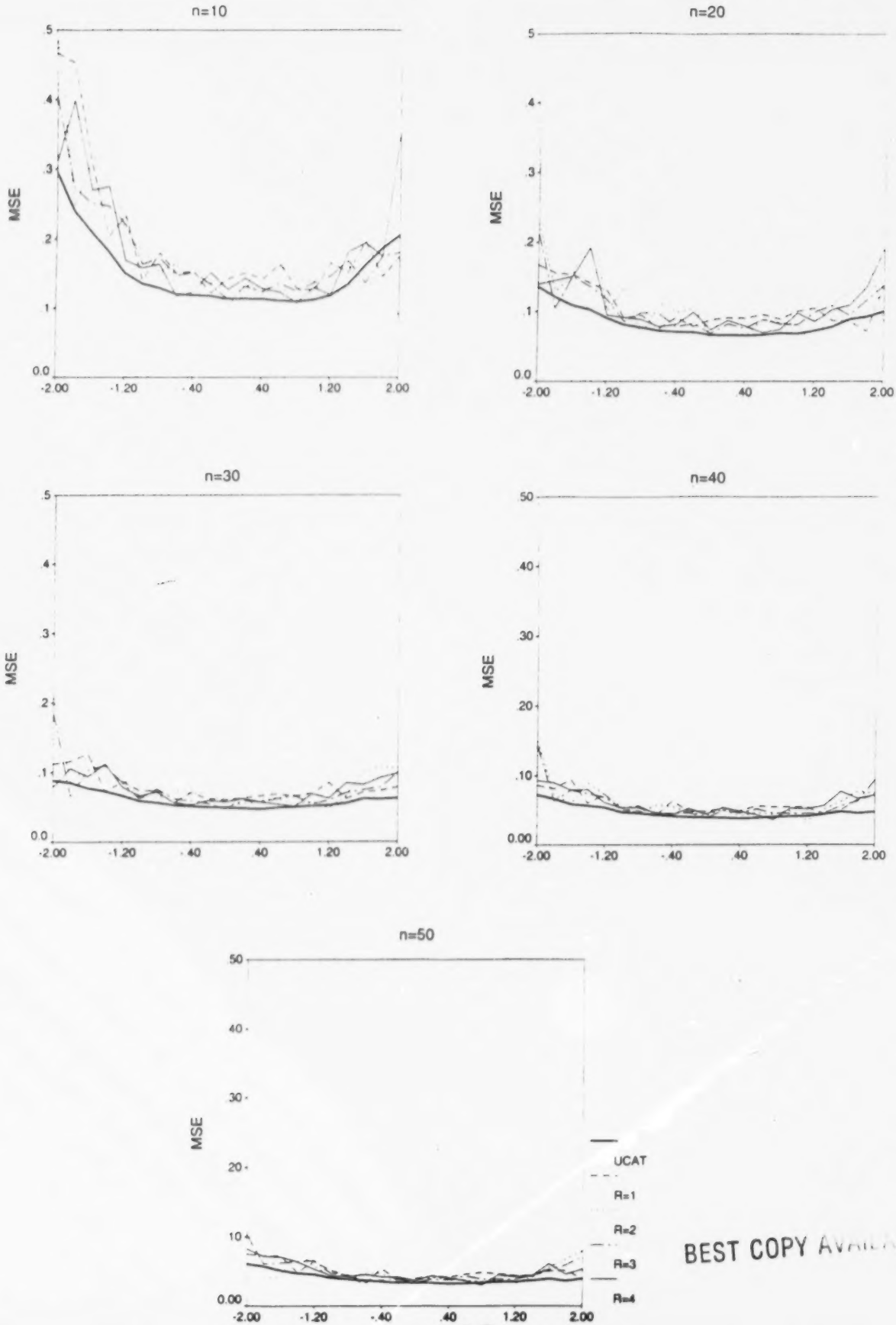
BEST COPY AVAILABLE



BEST COPY AVAILABLE



BEST COPY AVAILABLE



BEST COPY AVAILABLE

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-99-02 W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*

RR-99-01 R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*

RR-98-16 J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*

RR-98-15 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*

RR-98-14 A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*

RR-98-13 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Person Fit based on Statistical Process Control in an AdaptiveTesting Environment*

RR-98-12 W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*

RR-98-11 W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*

RR-98-10 W.J. van der Linden, *A Discussion of Some Methodological Issues in International Assessments*

RR-98-09 B.P. Veldkamp, *Multiple Objective Test Assembly Problems*

RR-98-08 B.P. Veldkamp, *Multidimensional Test Assembly Based on Lagrangian Relaxation Techniques*

RR-98-07 W.J. van der Linden & C.A W. Glas, *Capitalization on Item Calibration Error in Adaptive Testing*

RR-98-06 W.J. van der Linden, D.J. Scrams & D.L.Schnipke, *Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing*

RR-98-05 W.J. van der Linden, *Optimal Assembly of Educational and Psychological Tests, with a Bibliography*

RR-98-04 C.A.W. Glas, *Modification Indices for the 2-PL and the Nominal Response Model*

RR-98-03 C.A.W. Glas, *Quality Control of On-line Calibration in Computerized Assessment*

RR-98-02 R.R. Meijer & E.M.L.A. van Krimpen-Stoop, *Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests*

RR-98-01 C.A.W. Glas, R.R. Meijer, E.M.L.A. van Krimpen-Stoop, *Statistical Tests for Person Misfit in Computerized Adaptive Testing*

RR-97-07 H.J. Vos, *A Minimax Sequential Procedure in the Context of Computerized Adaptive Mastery Testing*

RR-97-06 H.J. Vos, *Applications of Bayesian Decision Theory to Sequential Mastery Testing*

RR-97-05 W.J. van der Linden & Richard M. Luecht, *Observed-Score Equating as a Test Assembly Problem*

RR-97-04 W.J. van der Linden & J.J. Adema, *Simultaneous Assembly of Multiple Test Forms*

RR-97-03 W.J. van der Linden, *Multidimensional Adaptive Testing with a Minimum Error-Variance Criterion*

RR-97-02 W.J. van der Linden, *A Procedure for Empirical Initialization of Adaptive Testing Algorithms*

RR-97-01 W.J. van der Linden & Lynda M. Reese, *A Model for Optimal Constrained Adaptive Testing*

RR-96-04 C.A.W. Glas & A.A. Béguin, *Appropriateness of IRT Observed Score Equating*

RR-96-03 C.A.W. Glas, *Testing the Generalized Partial Credit Model*

RR-96-02 C.A.W. Glas, *Detection of Differential Item Functioning using Lagrange Multiplier Tests*

RR-96-01 W.J. van der Linden, *Bayesian Item Selection Criteria for Adaptive Testing*

RR-95-03 W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*

RR-95-02 W.J. van der Linden, *Stochastic Order in Dichotomous Item Response Models for Fixed Tests, Adaptive Tests, or Multiple Abilities*

RR-95-01 W.J. van der Linden, *Some decision theory for course placement*

RR-94-17 H.J. Vos, *A compensatory model for simultaneously setting cutting scores for selection-placement-mastery decisions*

RR-94-16 H.J. Vos, *Applications of Bayesian decision theory to intelligent tutoring systems*

RR-94-15 H.J. Vos, *An intelligent tutoring system for classifying students into Instructional treatments with mastery scores*

RR-94-13 W.J.J. Veerkamp & M.P.F. Berger, *A simple and fast item selection procedure for adaptive testing*

RR-94-12 R.R. Meijer, *Nonparametric and group-based person-fit statistics: A validity study and an empirical example*

...

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, Mr. J.M.J. Nelissen, P.O. Box 217, 7500 AE Enschede, The Netherlands.

faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

BEST COPY AVAILABLE

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Ens[redacted]